Journal of Biomedical Informatics 54 (2015) 329–336

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs

Bridget T. McInnes [a,*], Ted Pedersen [b]

[a] Department of Computer Science, Virginia Commonwealth University, 401 S. Main St., Rm E4225, Richmond, VA 23284, USA
[b] Department of Computer Science, University of Minnesota, 1114 Kirby Drive, Duluth, MN 55812, USA

ARTICLE INFO

Article history:
Received 4 June 2014
Accepted 13 November 2014
Available online 15 December 2014

Keywords:
Natural language processing
NLP
Semantic similarity
Semantic relatedness

ABSTRACT

Introduction: This article explores how measures of semantic similarity and relatedness are impacted by the semantic groups to which the concepts they are measuring belong. Our goal is to determine if there are distinctions between homogeneous comparisons (where both concepts belong to the same group) and heterogeneous ones (where the concepts are in different groups). Our hypothesis is that the similarity measures will be significantly affected since they rely on hierarchical *is-a* relations, whereas relatedness measures should be less impacted since they utilize a wider range of relations. In addition, we also evaluate the effect of combining different measures of similarity and relatedness. Our hypothesis is that these combined measures will more closely correlate with human judgment, since they better reflect the rich variety of information humans use when assessing similarity and relatedness.

Method: We evaluate our method on four reference standards. Three of the reference standards were annotated by human judges for relatedness and one was annotated for similarity.

Results: We found significant differences in the correlation of semantic similarity and relatedness measures with human judgment, depending on which semantic groups were involved. We also found that combining a definition based relatedness measure with an information content similarity measure resulted in significant improvements in correlation over individual measures.

Availability: The semantic similarity and relatedness package is an open source program available from http://umls-similarity.sourceforge.net/. The reference standards are available at http://www.people.vcu.edu/~{}btmcinnes/downloads.html.

1. Introduction

Semantic similarity and relatedness measures quantify the degree to which two concepts are similar (e.g., *liver-organ*) or related (e.g., *headache- aspirin*). Relatedness encompasses many kinds of relations, but generally shows how associated two concepts are with each other. For example, a headache can be *treated* with aspirin. Similarity is a specific relation that is a subset of relatedness, and is based on the degree to which two concepts are connected through hierarchical *is-a* relations. For example, *organ* could be an ancestor of *liver* in an *is-a* hierarchy, and would therefore have a high similarity score. *Headache* and *aspirin*, on the other hand, are not closely connected by any *is-a* relations, and so would have a low similarity score. However, since they may be connected by other kinds of relations (e.g., *treated by*) they could have a very high relatedness score.

The automated discovery of groups of semantically similar or related concepts and terms is critical to improving the retrieval [1] and clustering [2] of biomedical and clinical documents, and the development of biomedical terminologies and ontologies [3]. As such, a number of different similarity measures have been developed for the biomedical domain. These have been evaluated intrinsically via comparisons to various human reference standards [4,5], as well as extrinsically depending on how well they contribute to the performance of secondary applications [6,7]. However, to date there has been little work that considers the type of concept being evaluated. Our objective is to evaluate how measures of similarity and relatedness perform depending on the semantic groups of the concepts involved.

Similarity measures find paths between concepts in an *is-a* hierarchy. Concept pairs from different semantic groups may well be in different hierarchies and therefore not be connected by *is-a* relations. In addition, these different hierarchies may have different levels of granularity and coverage. Given these considerations, our hypothesis is that there will be a large degree of change in the correlation of similarity measures with human reference

* Corresponding author.
E-mail addresses: btmcinnes@vcu.edu (B.T. McInnes), tpederse@d.umn.edu (T. Pedersen).

http://dx.doi.org/10.1016/j.jbi.2014.11.014
1532-0464/

standards when the concepts in a pair are from different semantic groups. Our results support this hypothesis. We found that no single measure performed best over all the different semantic group pairs.

In this work, we also combined measures based on the hypothesis that measures of similarity and relatedness will be complementary, and may result in more robust measures that more closely correlate with human judgments. Our goal is to identify pairs of measures that provide complementary information that will improve our ability to quantify the degree of similarity and relatedness between two terms. Bill et al. [8] showed that a linear combination of the similarity measures proposed by Resnik [9] and Lin [10] increased the accuracy of identifying similar terms. The results, here in this paper, show that combining relatedness and similarity measures improved correlation scores overall. However, these results varied depending on the reference standard used and so no single pair of measures was found to always improve correlation.

This article is organized as follows. Section 2 provides an overview of the Unified Medical Language System (UMLS), which is our main source of data on concepts and their relations. Section 3 reviews the measures of semantic similarity and relatedness used in this study. Section 4 describes resources used beyond the UMLS for formulating some of the measures. The reference standards used in our evaluation are introduced in Section 5, and the details of our experiments on these standards are summarized in Section 6. Our results are presented in Section 7, and the article closes with our conclusions in Section 8.

2. Unified Medical Language System

The UMLS is a data warehouse containing three knowledge sources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus contains approximately 2 million biomedical and clinical concepts from over 100 different terminologies that have been semi-automatically integrated into a single source. One such source is the *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT), which is a comprehensive clinical terminology created for the electronic representation of clinical health information. The concepts in SNOMED CT are organized in a hierarchical structure in order to permit searching at various levels of specificity. The concepts are connected by two main types of hierarchical relations: *parent/child* (PAR/CHD) and *broader/narrower* (RB/RN). The PAR/CHD relations are strictly *is-a* relations while the RB/RN relations contain *part-of* relations.

The Semantic Network consists of a set of broad subject categories called semantic types in which each concept in the Metathesaurus is assigned one or more semantic type. For example, the semantic type of C0206250 [Autonomic nerve] is *Body Part, Organ, or Organ Component*. Currently, there exist 135 semantic types in the Semantic Network.

The SPECIALIST Lexicon contains terms that are used in the biomedical and health-related domain along with linguistic information such as spelling variants.

Included in the UMLS is also a categorization of semantic types referred to as *semantic groups*. A semantic group is a coarse grained grouping of the semantic types in the UMLS developed by [11] to provide a coarse-grained distinction between UMLS concepts based on their semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. Examples of semantic groups include: Anatomy, Phenomena, Disorders and Chemicals & Drugs. There currently exists 15 semantic groups.[1] Each CUI in the UMLS can be categorized by their semantic group.

[1] http://semanticnetwork.nlm.nih.gov/SemGroups/.

3. Similarity and relatedness measures

This section describes the similarity and relatedness measures used in this work.

3.1. Similarity measures

We classify the similarity measures into two broad categories: path-based and information content (IC)-based. The path-based similarity measures provide information about the co-location of the terms in a taxonomy. The IC measures use the taxonomy information but also include additional information about the concept with respect to its relationship with the other concepts. There are two methods used to calculate IC: *corpus-based* which uses the probability of the concept occurring in an external corpus, and *intrinsic-based* which uses the informativeness of a concept based on its placement within the taxonomy. The remainder of this subsection describes the various measures and how they are calculated.

3.1.1. Path-based measures

Rada et al. [1] introduce the Conceptual Distance measure, which is the length of the shortest path between two concepts ($c1$ and $c2$) in MeSH using RB/RN relations. Caviedes and Cimino [12] later evaluated this measure using the PAR/CHD relations. The *path* measure is a modification of this and is calculated as the reciprocal of the length of the shortest path as defined in Eq. (1).

$$\text{sim}_{path} = \frac{1}{\text{spath}(c_1, c_2)} \tag{1}$$

Wu and Palmer [13] extend this measure by incorporating the depth of the Least Common Subsumer (LCS). The LCS is the most specific ancestor two concepts share. In this measure, the similarity is twice the depth of the two concepts' LCS divided by the product of the depths of the individual concepts as defined in Eq. (2).

$$\text{sim}_{wup} = \frac{2 * \text{depth}(\text{lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \tag{2}$$

Leacock and Chodorow [14] extend the path measure by incorporating the depth of the taxonomy. Here, the similarity is the negative log of the shortest path (*spath*) between two concepts divided by twice the total depth of the taxonomy (D) as defined in Eq. (3).

$$\text{sim}_{lch} = -\log \frac{\text{spath}(c_1, c_2)}{2 * D} \tag{3}$$

3.1.2. Information Content (IC) measures

Information content (IC) is formally defined as the negative log of the probability of a concept. Resnik [9] modified IC to be used as a similarity measure. He defined the similarity of two concepts to be the IC of their LCS as shown in Eq. (4).

$$\text{sim}_{res} = \text{IC}(\text{lcs}(c_1, c_2)) = -\log(P(\text{lcs}(c_1, c_2))) \tag{4}$$

Jiang and Conrath [15] and Lin [10] extended Resnik's IC measure by incorporating the IC of the individual concepts. Lin defined the similarity between two concepts by taking the quotient between twice the IC of the concepts' LCS and the sum of the IC of the two concepts as shown in Eq. (5). This is similar to the measure proposed by Wu & Palmer; differing in the use of IC rather than the depth of the concepts.

$$\text{sim}_{lin} = \frac{2 * \text{IC}(\text{lcs}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)} \tag{5}$$

Jiang and Conrath defined the distance between two concepts to be the sum of the IC of the two concepts minus twice the IC of the concepts' LCS. We modify this measure to return a similarity score by taking the reciprocal of the distance as shown in Eq. (6).

$$\text{sim}_{jcn} = \frac{1}{\text{IC}(c_1) + \text{IC}(c_2) - 2 * \text{IC}(\text{lcs}(c_1, c_2))} \quad (6)$$

3.2. Information content

The information content of a concept can be calculated using information derived from a corpus (corpus-based) or information derived from a taxonomy (intrinsic-based). In this section, we describe both techniques.

As previously stated, IC is defined as the negative log of the probability of a concept. For corpus IC, we calculate the probability of a concept, c, by summing the probability of the concept, $P(c)$, occurring in some text plus the probability its descendants, $P(d)$, occurring in the same text as seen in Eq. (7).

$$P(c*) = P(c) + \sum_{d \in descendant(c)} P(d) \quad (7)$$

The initial probability of a concept, $P(c)$, and its descendants, $P(d)$, is obtained by dividing the number of times a concept is seen in the corpus, $freq(d)$, by the total number of concepts, N, in the corpus as seen in Eq. (8).

$$P(d) = freq(d)/N \quad (8)$$

The challenge with probability calculations for concepts is that a large number of annotations are required in order to provide sufficient coverage of the underlying taxonomy to achieve reasonable estimates. Intrinsic IC seeks to alleviate this problem while still capturing the generality and concreteness of a concept. It assess the informativeness of concept based on its placement within the hierarchy by looking at its incoming (ancestors) and outgoing (descendant) links.

In this work, we use the intrinsic IC calculation proposed by Sanchez et al. [16] defined in Eq. (9).

$$IC(c) = -log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|} + 1}{max\_leaves + 1}\right) \quad (9)$$

where *leaves* are the number of descendants of concept c that are leaf nodes, *subsumers* are the number of concept c's ancestors and *max_leaves* are the total number of leaf nodes in the taxonomy.

3.3. Relatedness measures

Lesk [17] introduces a measure that determines the relatedness between two concepts by counting the number of overlaps between their two definitions. An overlap is the longest sequence of one or more consecutive terms that occur in both definitions. When implementing this measure in WordNet, Banerjee and Pedersen [18] found that the definitions were short, and did not contain enough overlaps to distinguish between multiple concepts, therefore, they extended this measure by including the definitions of the related concepts.

Patwardhan and Pedersen [19] extend this measure using second-order co-occurrence vectors. In this method, a vector is created for each word in the concept's definition containing terms that co-occur with it in a corpus. These word vectors are averaged to create a single co-occurrence vector for the concept. The similarity between the concepts is calculated by taking the cosine between the concepts' second-order vectors. Liu et al. [5] modify and extend this measure to be used to quantify the relatedness between biomedical and clinical terms in the UMLS.

4. Resources

This section describes the data sources used in formulating the information content and vector measures. These sources are not a part of the UMLS and so are discussed here separately.

4.1. IC similarity measure data

The IC similarity measure data is used to calculate the probability of a concept occurring in a corpus. We use the *UMLSonMedline* dataset created by NLM which consists of concepts from the 2009AB UMLS and the number of times they occurred in a snapshot of Medline taken on 12 January, 2009. The frequency counts were obtained by using the Essie Search Engine proposed by Ide et al. [20] which queried Medline with normalized strings from the 2009AB MRCONSO table in the UMLS. The frequency of a CUI was obtained by aggregating the frequency counts of the terms associated with the CUI to provide a rough estimate of its frequency. The IC measures use this information to calculate the probability of a concept.

4.2. Relatedness measure data

The relatedness measure data is used by the *vector* measure to build the second-order co-occurrence matrix. We use the vector matrices, developed by Liu et al. [5], that are included in the UMLS::Similarity package. These matrices were created using the inpatient clinical reports that were collected from 2003 to 2008 at Fairview Health Services. These semi-structured reports consist of admission history, physical operation, discharge summaries, and consultation notes. These reports contain on average 500 words; after pre-processing (e.g. removal of stop words, numerals and punctuation), each note contained approximately 300 words. The total size of the resulting reports consisted of approximately 208.7 million words.

5. Reference standards

We use four reference standards[2] to evaluate the semantic similarity and relatedness measures: the UMNSRS tagged for similarity, the UMNSRS tagged for relatedness, the MayoSRS tagged for relatedness and the MiniMayoSRS tagged for relatedness. In this section, we describe the reference standards and then briefly discuss some of their differences.

5.1. MayoSRS

MayoSRS, developed by Pakhomov et al. [21], consists of 101 clinical term pairs whose relatedness was determined by nine medical coders and three physicians from the Mayo Clinic. The relatedness of each term pair was assessed based on a four point scale: (4.0) practically synonymous, (3.0) related, (2.0) marginally related and (1.0) unrelated. We evaluate our method on the mean score of the physicians and medical coders as provided by Pakhomov et al. [21].

5.2. MiniMayoSRS

MiniMayoSRS is a subset of the MayoSRS and consists of 30 term pairs on which a higher inter-annotator agreement was achieved. The average correlation between physicians is 0.68. The average correlation between medical coders is 0.78. We evaluate our method on the mean of the physician scores and the mean of

[2] http://www.people.vcu.edu/~btmcinnes/downloads.html.

Table 1
Semantic groupings of term pairs in the reference standards.

Semantic group		MayoSRS	MiniMayoSRS	UMNSRS	
Term 1	Term 2	Rel.	Rel.	Sim.	Rel.
Activities & Behaviors	Phenomena	1			
Anatomy	Anatomy	1	1		
Chemical & Drug	**Chemical & Drug**			**77**	**82**
Chemical & Drug	Devices	1			
Chemical & Drug	Procedures	1	1		
Disorder	Anatomy	4	2		
Disorder	**Chemical & Drug**	**10**	**1**	**113**	**126**
Disorder	Concepts & Ideas	3	1		
Disorder	**Disorder**	**66**	**21**	**211**	**222**
Disorder	Devices		1		
Disorder	Physiology	5			
Disorder	Procedures	7	1		
Physiology	Physiology	1			
Total		100	29	401	430

The bold values referred to the term pairs in the MiniMayoSRS and MayoSRS consist primarily of Disorders.

the coders' scores in this subset in the same manner as reported by Pedersen et al. [22].

5.3. UMNSRS

UMNSRS, developed by Pakhomov et al. [23], consists of 725 clinical term pairs whose semantic similarity and relatedness was determined independently by four medical residents from the University of Minnesota Medical School. The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness. The Intraclass Correlation Coefficient (ICC) for the reference standard tagged for similarity was 0.47, and 0.50 for relatedness. Therefore, as suggested by Pakhomov and colleagues,we use a subset of the ratings consisting of 401 pairs for the similarity set and 430 pairs for the relatedness set which each have an ICC equal to 0.73.

5.4. Semantic group breakdown of datasets

As stated in Section 2, a semantic group is a coarse grained grouping of the semantic types. Each CUI in the UMLS can be categorized by their semantic group. Table 1 shows a breakdown of the semantic groups for the concepts in each of the reference standards.

6. Experimental framework

We conducted our experiments using the freely available open source software package UMLS::Similarity [24] version 1.13.[3] This package takes as input two terms or concepts and returns the similarity between any two concepts using the path information in any of the sources available in the UMLS, including SNOMED CT, for each of the measures discussed in Section 3.

These experiments were conducted using the 2013AB version of the UMLS. We use the SNOMED CT taxonomy located in the UMLS Metathesaurus for the similarity measures and the entire UMLS (Level 1 + SNOMED CT) for the relatedness measures. Correlation between the results of the similarity measures and human judgments were conducted using Spearman's Rank Correlation (ρ). Spearman's measures the statistical dependence between two variables to assess how well the relationship between the rankings of the variables can be described using a monotonic function. We used Fisher's r-to-z transformation [25] to calculate the significance between the correlation results.

[3] http://search.cpan.org/dist/UMLS-Similarity/.

As previously stated, the goal of combining the measures is to capitalize on each of the measures strengths in hopes that their combination provides complementary information for quantifying the degree of similarity/relatedness between the two terms. The path-based similarity measures provide information about the co-location of the terms in a taxonomy; the intrinsic IC measures provide information about the concept in relation to the other concepts in the taxonomy; the corpus IC measures provide probability information regarding its occurrence in an external corpus; and relatedness measures provide contextual information about the term. We combined the similarity and relatedness measure by first standardizing the individual scores to place them on the same scale and then averaging the standardized scores. We standardize the scores by subtracting the score by the sample mean and then dividing it by the standard deviation as shown in Eq. (10) where $\overline{score}$ is the mean and $stddev(score)$ is the standard deviation.

$$\text{standardized}(\text{score}) = \frac{score - \overline{score}}{\text{stddev}(score)} \tag{10}$$

7. Results and discussion

Table 2 shows the Spearman's Rank Correlation between the human scores from the four reference standards and the scores obtained by the path measure (*path*); the path-based measures proposed by Leacock & Chodorow [14] (*lch*) and Wu & Palmer [13] (*wup*); the IC measures proposed by Resnik [9] (*res*/*i-res*), Jiang & Conrath [15] (*jcn*/*i-jcn*) and Lin [10] (*lin*/*i-lin*) using the corpus and intrinsic IC respectively; the relatedness measures proposed by Lesk [17] (*lesk*) and Patwardhan & Pedersen [19] (*vector*); and the random baseline (*random*) which randomly assigns a term pair a similarity score between zero and one.

The results show that for the UMNSRS, the similarity measure *path*/*lch* obtained the highest correlation on the standard tagged for similarity ($p \leqslant 0.005$) and *lesk* obtained the highest correlation on the standard tagged for relatedness ($p \leqslant 0.03$). For the MayoSRS, *lesk* obtained the highest correlation ($p \leqslant 0.05$). For the MiniMayoSRS, the similarity measure *lin* obtained the highest correlation with the coders ($p \leqslant 0.02$) but the relatedness measure *lesk* obtained the highest correlation with the physicians ($p \leqslant 0.08$).

7.1. Semantic group results

In this section, we analyze how the measures perform over the semantic grouping associated with the term pairs. Table 3 shows a breakdown of the correlation results based on the term pairs' semantic groups in the UMNSRS reference standard tagged for

Table 2
Spearman's rank correlation results.

Measure	Reference standard				
	MiniMayoSRS		MayoSRS	UMNSRS	
	Coders	Physicians	Rel.	Sim.	Rel.
path	0.4488	0.3473	0.1857	**0.5335**	0.3062
wup	0.5088	0.3891	0.2024	0.5079	0.2633
lch	0.4488	0.3473	0.1857	**0.5335**	0.3062
res	0.4747	0.3632	0.2549	0.4905	0.2800
jcn	0.5168	0.4239	0.3199	0.5267	0.3555
lin	**0.5396**	0.4405	0.3051	0.5137	0.3114
i-res	0.4950	0.3962	0.2580	0.4910	0.2797
i-jcn	0.4927	0.3874	0.3195	0.5154	0.3081
i-lin	0.5031	0.4065	0.2813	0.5052	0.3023
lesk	0.5044	**0.4965**	**0.3676**	0.5246	**0.4379**
vector	0.4372	0.4192	0.2902	0.5289	0.4048
random	0.1488	−0.0280	−0.0184	−0.0027	−0.0455

Table 3
Semantic group breakdown of UMNSRS tagged for relatedness.

Measure	D–D	D–C	C–C	Overall
path	0.4746	−0.1138	0.3859	0.3062
wup	0.4307	−0.1138	0.4196	0.2633
lch	0.4746	−0.1138	0.3859	0.3062
res	0.4083	n/a	0.3955	0.2800
jcn	**0.4805**	0.1249	0.4103	0.3555
lin	0.4774	0.1295	0.4120	0.3114
i-res	0.4141	n/a	0.3987	0.2797
i-jcn	0.4447	0.1403	0.3692	0.3081
i-lin	0.4446	0.1418	0.3685	0.3023
lesk	0.4336	**0.3463**	**0.4515**	**0.4379**
vector	0.4335	0.2582	0.4099	0.4048

Table 4
Semantic group breakdown of UMNSRS tagged for similarity.

Measure	D–D	D–C	C–C	Overall
path	**0.5380**	0.1839	0.5963	**0.5335**
wup	0.4839	0.1839	0.6796	0.5079
lch	**0.5380**	0.1839	0.5963	**0.5335**
res	0.4284	0.0000	0.6990	0.4905
jcn	0.4709	0.1888	0.6855	0.5267
lin	0.4894	0.1966	0.6932	0.5137
i-res	0.4340	0.0000	0.7188	0.4910
i-jcn	0.4810	0.1266	0.7088	0.5154
i-lin	0.4628	0.1303	0.7171	0.5052
lesk	0.4297	**0.3148**	**0.7319**	0.5246
vector	0.4863	0.2175	0.6669	0.5289

relatedness and the overall correlation score for reference; Table 4 shows the same results on the reference standard tagged for similarity; and Table 5 shows the results for the MiniMayoSRS and MayoSRS references standards.

As previously discussed, the semantic groups of the terms in the reference standards can be classified as Disorder–Disorder (D–D), Disorder–Chemical&Drug (D–C), and Chemical&Drug–Chemical&Drug (C–C).

The UMNSRS tagged for relatedness results in Table 3 show that for the D–D pairs, the corpus IC measure *jcn* obtained the highest correlation ($p \leqslant 0.003$). The results indicate that the information captured within the taxonomy is a better indicator of two Disorder's (D–D) relatedness than the contextual information used by the relatedness measures. This is not the case for the C–C pairs where the relatedness measure *lesk* obtained the highest correlation ($p \leqslant 0.03$); indicating that unlike Disorders terms the contextual information describing Chemical&Drug terms provides a better indicator of their relatedness than their location within

Table 5
Semantic group breakdown of MiniMayoSRS and MayoSRS.

Measure	MiniMayoSRS				MayoSRS	
	Coders		Physicians			
	D–D	Overall	D–D	Overall	D–D	Overall
path	0.2228	0.4488	0.1686	0.3473	0.2345	0.1857
wup	0.3012	0.5088	0.2336	0.3891	0.2265	0.2024
lch	0.2228	0.4488	0.1686	0.3473	0.2346	0.1857
res	0.2861	0.4747	0.2287	0.3632	0.2735	0.2549
jcn	0.2697	0.5168	0.2288	0.4239	0.2770	0.3199
lin	0.2959	**0.5396**	0.2494	0.4405	0.3018	0.3051
i-res	0.3070	0.4950	0.2574	0.3962	0.2619	0.2580
i-jcn	0.2842	0.4927	0.2279	0.3874	0.3561	0.3195
i-lin	0.3004	0.5031	0.2601	0.4095	0.2876	0.2813
lesk	**0.5106**	0.5044	**0.5730**	**0.4965**	**0.4409**	**0.3676**
vector	0.4510	0.4372	0.4917	0.4192	0.3195	0.2902

the taxonomy. The D–C pairs results are not surprising given that there is little taxonomic information that connects a Disorder term and a Chemical&Drug term through an *is-a* hierarchy. The results show that the relatedness measures *lesk* and *vector* obtained the highest correlation ($p \leqslant 0.1$).

The results for the UMNSRS reference standard tagged for similarity shown in Table 4 are consistent with Table 3, although the correlation for the homogeneous D–D and C–C pairs is considerably higher. The results for the D–D pairs shows that the similarity measures path and *lch* have the highest correlation ($p \leqslant 0.05$). It is perhaps not surprising that a reference sample where humans have specifically rated for similarity should result in high correlation with similarity measures, although it is important to see such intuitions confirmed experimentally.

However, for the C–C pairs there is again a slightly counter-intuitive result (as was also seen in Table 3) where the *lesk* relatedness measure performs at a significantly ($p \leqslant 0.01$) higher level of correlation (0.7319). This is perhaps surprising since with the pairs rated for relatedness the *lesk* measure attained a somewhat lower correlation (0.4519 from Table 3). One possible intuition to explain these results is that similar concepts should have similar definitions, and that *lesk* is able to capture this particularly well. Why this would be significantly better than the information obtained from positions in the taxonomy remains an interesting question for future work.

The term pairs in the MiniMayoSRS and MayoSRS consist primarily of Disorders (D–D); the number of additional term pairs in each category are too small to evaluate as shown in Table 1. The results show that the relatedness measures obtained a higher correlation than similarity measures between the Disorder pairs (D–D) with *lesk* obtaining the highest ($p \leqslant 0.08$). This is contrary to what we saw in Tables 3,4 for the UMNSRS reference standards where similarity measures attained the highest correlation for the D–D pairs.

7.2. Combination results

Table 6 shows the results when combining each of the similarity and relatedness measures on the UMNSRS. The upper right portion of the table are the results using the reference standard tagged for relatedness, and the lower left are the results using the reference standard tagged for similarity.

The results show that for the UMNSRS tagged for relatedness the combination of the *i-jcn* and the *lesk* measures obtain the highest correlation with human judgments (0.4664). This is significantly higher ($p \leqslant 0.03$) than the highest individual results shown previously in Table 2 where *lesk* obtained the highest correlation (0.4379). One might expect that the combination of *lesk* and

Table 6
Combination results for UMNSRS tagged for relatedness (upper) and similarity (lower).

	path	wup	lch	lin	res	jcn	ilin	ires	ijcn	lesk	vector
path		0.2778	0.3059	0.2969	0.2830	0.3198	0.2909	0.2847	0.3100	0.3858	0.4418
wup	0.5158		0.2822	0.2824	0.2731	0.2801	0.2809	0.2740	0.2745	0.3394	0.3892
lch	0.5342	0.5192		0.3027	0.2867	0.3192	0.2979	0.2901	0.3099	0.3669	0.4230
lin	0.5271	0.5225	0.5334		0.3055	0.3099	0.3070	0.3038	0.3088	0.3670	0.4106
res	0.5202	0.5132	0.5252	0.5120		0.3037	0.2985	0.2794	0.2981	0.3600	0.4097
jcn	0.5476	0.5167	0.5470	0.5137	0.5073		0.3078	0.3033	0.3417	0.4658	0.4413
ilin	0.5261	0.5233	0.5326	0.5086	0.5035	0.5071		0.2998	0.3024	0.3665	0.4122
ires	0.5199	0.5150	0.5265	0.5108	0.4910	0.5062	0.5039		0.2976	0.3619	0.4111
ijcn	0.5432	0.5181	0.5438	0.5140	0.5068	0.5315	0.5052	0.5065		**0.4664**	0.4445
lesk	0.5615	0.5385	0.5601	0.5364	0.5363	0.5657	0.5356	0.5397	0.5770		0.4365
vector	0.5901	0.5724	**0.5923**	0.5746	0.5768	0.5660	0.5802	0.5790	0.5787	0.5394	

path would produce the higher correlation, since these two measures are very different from each other and seem likely candidates to be complementary. However, in this case the correlation of *lesk* on its own is actually much better (0.4379, see Table 2) than the combination of *lesk* and *path* (0.3858, see Table 6).

For the UMNSRS tagged for similarity, the combination of the *lch* and *vector* measures obtains the highest correlation (0.5923). This is significantly higher ($p \leqslant 0.05$) than the highest individual results where *path*/*lch* obtained the highest correlation (0.5335).

Table 7 shows the results when combining each of the similarity and relatedness measures on the MiniMayoSRS. The upper right portion of the table are the coders' correlation results and the lower left are the physician's correlation results. These results show the *i-jcn* and the *lesk* measures obtain the highest correlation for both datasets. As with the UMNSRS results, the combination of the relatedness and similarity measure obtained a higher correlation with the human judgments. For the set tagged by coders, the combined correlation results (0.7352) are higher than the highest individual results obtained by *lin* (0.5396) but given the sample size the results are not significant ($p \leqslant 0.19$). Similarly, for the set tagged by physicians, the combined correlation results (0.6753) are higher than the highest individual results obtained by *lesk* (0.4965) but the results are not significant ($p \leqslant 0.17$).

Table 8 shows the results when combining each of the similarity and relatedness measures on the MayoSRS. The results show that the *i-jcn* and *vector* measures obtain the highest correlation with human judgments (0.4206). This is significantly higher ($p \leqslant 0.05$) than the highest individual results shown previously in Table 2 where *lesk* obtained the highest correlation (0.3676).

To summarize the increase in results when combining the measures, Table 9 shows the highest correlation using the individual

Table 9
Summary of Spearman's rank best original and combination correlations results.

Reference standard		Original	Combination
UMNSRS	Rel.	0.4379	0.4664
	Sim.	0.5335	0.5923
MiniMayo	Physician	0.4965	0.6753
	Coders	0.5396	0.7352
Mayo	Rel.	0.3676	0.4206

Table 7
Combination results for MiniMayoSRS with coders (upper) and physicians (lower).

	path	wup	lch	lin	res	jcn	ilin	ires	ijcn	lesk	vector
path		0.4807	0.4488	0.4783	0.5016	0.5183	0.4974	0.5016	0.4790	0.5875	0.5370
wup	0.3799		0.4724	0.5226	0.5077	0.5161	0.5118	0.5095	0.4977	0.5406	0.5370
lch	0.3473	0.3775		0.5013	0.4969	0.4793	0.4963	0.4875	0.4624	0.5193	0.5053
lin	0.3756	0.4190	0.3971		0.5284	0.5309	0.5175	0.5411	0.5341	0.5608	0.5925
res	0.3723	0.3859	0.3730	0.4193		0.5276	0.5031	0.4958	0.5114	0.5968	0.5806
jcn	0.4257	0.4006	0.4010	0.4396	0.4152		0.5147	0.5276	0.5287	0.7345	0.5979
ilin	0.3811	0.4051	0.3867	0.4201	0.4090	0.4086		0.5024	0.5031	0.5622	0.5936
ires	0.3838	0.3970	0.3733	0.4370	0.3944	0.4277	0.4107		0.5031	0.5698	0.5950
ijcn	0.3919	0.3850	0.3704	0.4260	0.4197	0.4267	0.4065	0.4090		**0.7352**	0.6260
lesk	0.5411	0.4733	0.4618	0.4809	0.5268	0.6569	0.5095	0.4872	**0.6753**		0.4649
vector	0.4601	0.4733	0.4253	0.5140	0.5042	0.5303	0.5296	0.5237	0.5495	0.4414	

Table 8
Combination results for MayoSRS.

	path	wup	lch	lin	res	jcn	ilin	ires	ijcn	lesk	vector
path	0.1857	0.1861	0.1857	0.2533	0.2452	0.2543	0.2528	0.2440	0.2469	0.3799	0.3086
wup		0.2024	0.1811	0.2427	0.2482	0.2479	0.2435	0.2364	0.2433	0.3568	0.2759
lch			0.1857	0.2403	0.2352	0.2346	0.2389	0.2338	0.2436	0.3715	0.2864
lin				0.3051	0.2991	0.2960	0.2928	0.2970	0.3023	0.4037	0.3565
res					0.2516	0.3107	0.2730	0.2592	0.2878	0.3896	0.3816
jcn						0.3199	0.3047	0.3004	0.3371	0.4183	0.3477
ilin							0.2813	0.2753	0.2937	0.4062	0.3698
ires								0.2580	0.2843	0.3990	0.3692
ijcn									0.3195	0.4198	**0.4206**
lesk										0.3676	0.3667
vector											0.2902

measures and the highest correlation achieved when combining the measures for each of the reference standards. The results show that overall combining the measures obtains a higher correlation to human judgments.

7.3. Correlation results between measures

To analyze the individual measures, we calculated the correlation of the rankings between them. Table 10 shows the correlation using the UMNSRS tagged for relatedness (upper) and similarity (lower). Table 11 shows similar results for the MiniMayoSRS (upper) and MayoSRS (lower).

The correlation results show that for the UMNSRS, the combination of relatedness measures and path-based measures obtain a higher correlation with human judgments than the individual measures. However, with the MiniMayoSRS and MayoSRS results, it is the combination of relatedness measures and IC similarity measures that attain the highest correlation.

For the UMNSRS data tagged for relatedness, the combination of the *lesk* relatedness measure and the *i-jcn* (IC similarity measure) obtain the highest correlation with human judgments (0.4664, from Table 6). Note as well that in Table 10 the relatively low correlation between the measures (0.6172) provides some evidence of their complementary nature. However, it's important to point out that the measures with the least correlation between them (0.5205 from Table 10, for the path based *wup* measure and the *lesk* relatedness measure) end up having rather moderate correlation with human judgments (0.3394, from Table 6).

For the UMNSRS data tagged for similarity, the combination of the *vector* relatedness measure with either the *path* or *lch* path-based measures results in the highest correlation with human judgments (0.5923 from Table 6). Once again their between measure correlation is a relatively low value compared to many of the other pairs (0.5814 from Table 10).

For the MiniMayoSRS and MayoSRS, both of which were tagged for relatedness, the combination of the *lesk* relatedness measure and *i-jcn* (IC-similarity measure) obtained the highest correlation with human judgments (see Tables 7, 8) and also have the lowest between measure correlation for relatedness (0.4523 from Table 11).

These results suggest that finding pairs of measures that have low correlation between could identify measures that are complementary to each other, and that may attain higher correlations to human judgments than the individual measures. However, as our results show this is not always exactly the case, and so this between measure correlation (as shown in Tables 10, 11) should only be considered as one of several possible means of deciding how to combine measures. However, we do observe that the combination of relatedness measures with IC similarity measures may be better suited to quantify relatedness, whereas relatedness measures combined with path-based similarity measures may be better able to measure similarity. The methods by which effective combinations of measures may be discovered seems to be a particularly promising avenue for future research.

7.4. Combination results over semantic groups

In this work, we also analyzed how well combinations of measures performed on the different semantic groupings. Table 12 shows the combination results of the similarity measures *i-jcn* and *lch* combined with *lesk* and *vector*, and the highest correlation obtained by the individual measures (*indiv. meas.*) from Tables 3,4 for each of the reference standards and the semantic grouping pairs that have at least twenty term pairs in their respective datasets.

The results show that overall using *lesk* in combination with *i-jcn* attains the highest correlation for most of the homogeneous pairs (D–D and C–C), although there are a few cases where using *vector* or *lesk* with *lch* perform the best. In general, this shows that the homogeneous pairs can be significantly improved via measures that combined relatedness and similarity measures. The heterogeneous pairs (D–C) do not tend to improve as much (if at all) with combined measures.

Table 10
Correlation of measures for UMNSRS tagged for relatedness (upper) and similarity (lower).

	path	wup	lch	lin	res	jcn	ilin	ires	ijcn	lesk	vector
path		0.9499	1.0000	0.8726	0.8627	0.8189	0.8748	0.8693	0.8721	0.5275	0.6024
wup	0.9940		0.9499	0.9450	0.9500	0.8505	0.9513	0.9595	0.9383	**0.5205**	0.5945
lch	1.0000	0.9440		0.8726	0.8627	0.8189	0.8748	0.8693	0.8721	0.5275	0.6024
lin	0.8546	0.9413	0.8549		0.9671	0.9348	0.9779	0.9643	0.9711	0.6190	0.6360
res	0.8556	0.9532	0.8556	0.9696		0.8440	0.9709	0.9930	0.9510	0.5459	0.5726
jcn	0.7851	0.8360	0.7851	0.9338	0.8444		0.8886	0.8430	0.9056	0.6936	0.6919
ilin	0.8638	0.9514	0.8638	0.9814	0.9735	0.8882		0.9776	0.9922	0.5918	0.6255
ires	0.8594	0.9602	0.8594	0.9670	0.9955	0.8401	0.9777		0.9567	0.5323	0.5741
ijcn	0.8572	0.9335	0.8572	0.9719	0.9485	0.9067	0.9904	0.9522		0.6172	0.6452
lesk	**0.5419**	0.5568	**0.5419**	0.6467	0.5633	0.7283	0.6204	0.5579	0.6506		0.6453
vector	0.5814	0.5872	0.5814	0.6313	0.5625	0.6955	0.6166	0.5642	0.6405	0.7001	

Table 11
Correlation of measures for MiniMayoSRS (upper) and MayoSRS (lower).

	path	wup	lch	lin	res	jcn	ilin	ires	ijcn	lesk	vector
path		0.8990	1.0000	0.8728	0.7923	0.8531	0.8197	0.7712	0.8593	0.5446	0.6349
wup	0.9156		0.8990	0.9642	0.9431	0.9139	0.9683	0.9431	0.9749	0.4544	0.5779
lch	1.0000	0.9156		0.8728	0.7923	0.8531	0.8192	0.7712	0.8593	0.5446	0.6349
lin	0.8134	0.9224	0.8134		0.9490	0.9574	0.9732	0.9428	0.9855	0.5639	0.6603
res	0.7881	0.9316	0.7881	0.9603		0.8703	0.9830	0.9917	0.9506	0.4718	0.5232
jcn	0.7411	0.7761	0.7411	0.8980	0.7659		0.9207	0.8606	0.9453	0.5699	0.7470
ilin	0.8194	0.9446	0.8194	0.9751	0.9720	0.8195		0.9844	0.9802	0.4851	0.5624
ires	0.7902	0.9395	0.7902	0.9573	0.9911	0.7641	0.9799		0.9457	**0.4532**	0.4791
ijcn	0.8209	0.9248	0.8209	0.9678	0.9358	0.8566	0.9837	0.9445		0.4803	0.6362
lesk	0.4677	0.5138	0.4677	0.5825	0.4889	0.6809	0.5426	0.4844	0.5982		0.7256
vector	0.3813	0.4039	0.3813	0.4205	**0.3202**	0.5396	0.3835	0.3255	0.4377	0.7608	

Table 12
Combination results of the semantic groups.

Reference standard		SGs	Indiv. meas.	lch/lesk	lch/vector	i-jcn/lesk	i-jcn/vector
UMNSRS	Rel.	D–D	0.4805 (*jcn*)	0.5492	**0.5714**	0.5501	0.5354
		D–C	**0.3463** (*lesk*)	0.1529	0.0676	0.2955	0.2116
		C–C	0.4515 (*lesk*)	**0.4530**	0.4117	0.4233	0.3456
	Sim.	D–D	0.5380 (*lch*)	0.5629	**0.6143**	0.5346	0.5709
		D–C	0.3148 (*lesk*)	**0.3336**	0.2784	0.2375	0.1951
		C–C	0.7319 (*lesk*)	0.7016	0.6711	**0.7684**	0.6625
MiniMayo	Physician	D–D	0.5730 (*lesk*)	0.2655	0.2842	**0.6006**	0.4210
	Coders	D–D	0.5106 (*lesk*)	0.2589	0.2932	**0.5530**	0.3915
Mayo	Rel.	D–D	0.4409 (*lesk*)	0.5096	0.3115	**0.5242**	0.4840

In summary, the individual results discussed previously in Tables 3 and 4 and the combination results discussed here in Table 12 show that different measures attain higher correlations with human judgments, and that this appears to be dependent on the particular semantic groups involved in the semantic grouping pair. Similarity measures do not fare well in heterogeneous comparisons (D–C). While relatedness are somewhat more successful, there is still room for improvement.

8. Conclusions

In this paper, we analyzed the correlation of semantic similarity and relatedness measures to human judgments over various semantic groupings of the term pairs. The results show that no one measure performed best over all of the semantic grouping pairs analyzed in this article. The results also indicated that using similarity measures on term pairs across disparate semantic groupings does not result in high correlation because there is little taxonomy information connecting them; relatedness measures are a better choice for these term pairs. In the future, we would like to analyze additional semantic groupings in order to determine if this is the case across all semantic groups.

We also analyzed the results of combining various measures to determine if they are complementary. The results showed that combining relatedness and similarity measures improved the correlation scores; specifically we found that using *lesk* and the *i-jcn* measure obtained the highest overall correlation over the datasets; although *vector* with *lch* or *i-jcn* also performed well. In the future, we plan to explore and develop additional measures that incorporate aspects of semantic similarity and relatedness measures into a single measure.

In the future, we also plan to evaluate what constitutes a high correlation and at what level do the measures need to correlate with human judgments to be practically useful. In an intrinsic evaluation, such as this study, it is difficult to pick out a particular correlation level and say 'this is good enough'. Therefore, we plan to conduct an extrinsic evaluation by analyzing the measures with respect to their correlation in a secondary application.

Acknowledgments

We would like to thank Russel Loane, Jim Mork and Lan Aronson from NLM for providing the UMLSonMedline dataset.

References

[1] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybernet 1989;19(1):17–30.
[2] Lin Y, Li W, Chen K, Liu Y. A document clustering and ranking system for exploring MEDLINE citations. J Am Med Inform Assoc 2007;14(5):651–61.
[3] Bodenreider O, Burgun A. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. In: Proceedings of the 11th World Congress on Medical Informatics (MEDINFO), San Fransico, CA; 2004. p. 327–31.
[4] Workman E, Rosemblat G, Fiszman M, Rindflesch T. A literature-based assessment of concept pairs as a measure of semantic relatedness. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium, vol. 2013, American Medical Informatics Association; 2013. p. 1512–21.
[5] Liu Y, McInnes B, Pedersen T, Melton-Meaux G, Pakhomov S. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In: Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics. ACM; 2012. p. 363–72.
[6] McInnes BT, Pedersen T. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. J Biomed Inform 2013;46(6):1116–24.
[7] McInnes B, Stevenson M. Determining the difficulty of word sense disambiguation. J Biomed Inform 2014;47:83–90.
[8] Bill RW, Liu Y, McInnes BT, Melton GB, Pedersen T, Pakhomov S. Evaluating semantic relatedness and similarity measures with standardized MedDRA queries. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium, vol. 2012, American Medical Informatics Association; 2012. p. 43–50.
[9] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada; 1995. p. 448–53.
[10] Lin D. An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA; 1998. p. 296–304.
[11] McCray A, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 2001:216–20.
[12] Caviedes J, Cimino J. Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform 2004;37(2):77–85.
[13] Wu Z, Palmer M. Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics, Las Cruces, NM; 1994. p. 133–8.
[14] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: Fellbaum C, editor. WordNet: an electronic lexical database. Cambridge, MA: MIT Press; 1998. p. 265–83.
[15] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th Research on Computational Linguistics International Conference, Tapei, Taiwan; 1997. p. 19–33.
[16] Sánchez D, Batet M, Isern D. Ontology-based information content computation. Knowl-Based Syst 2011;24(2):297–303.
[17] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation, Toronto, Canada; 1986. p. 24–6.
[18] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico; 2003. p. 805–10.
[19] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 Workshop Making Sense of Sense – Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy; 2006. p. 1–8.
[20] Ide N, Loane R, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14(3):253–63.
[21] Pakhomov S, Pedersen T, McInnes B, Melton G, Ruggieri A, Chute C. Towards a framework for developing semantic relatedness reference standards. J Biomed Inform 2011;44(2):251–65.
[22] Pedersen T, Pakhomov S, Patwardhan S, Chute C. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40(3):288–99.
[23] Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton G. Semantic similarity and relatedness between clinical terms: An experimental study. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium, Washington, DC; 2010. p. 572–6.
[24] McInnes B, Pedersen T, Pakhomov S. UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium, San Fransico, CA; 2009. p. 431–5.
[25] Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 1915:507–21.